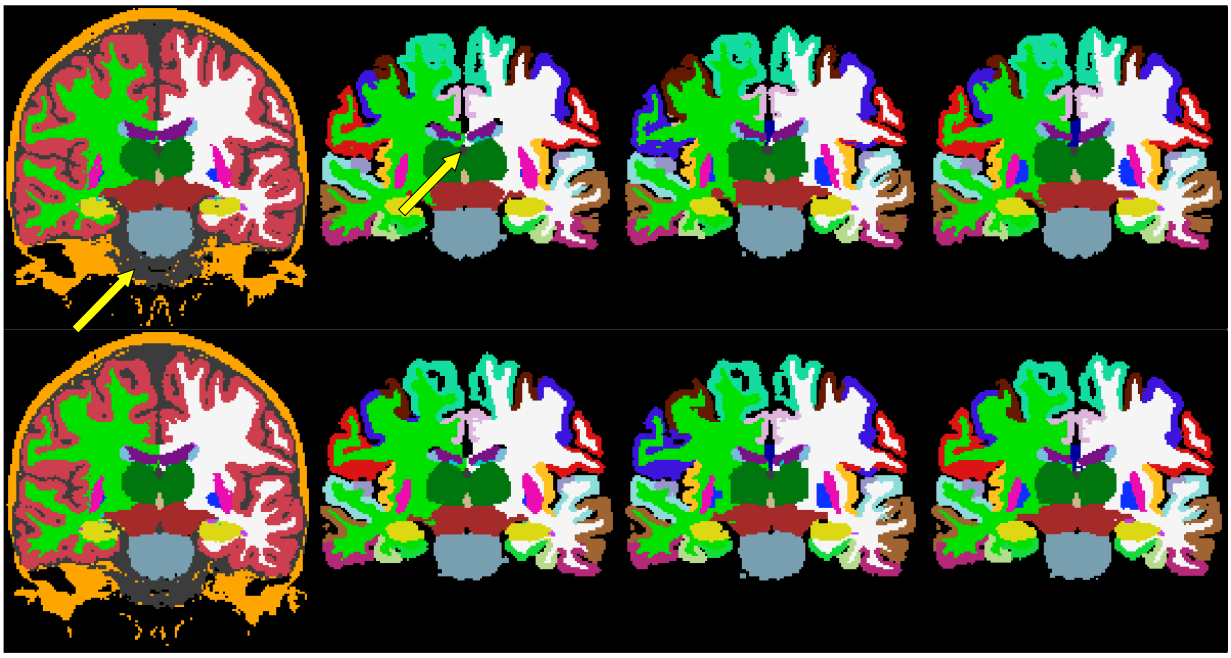
12. Supplement

Supplementary Figure S1. Examples of segmentation output

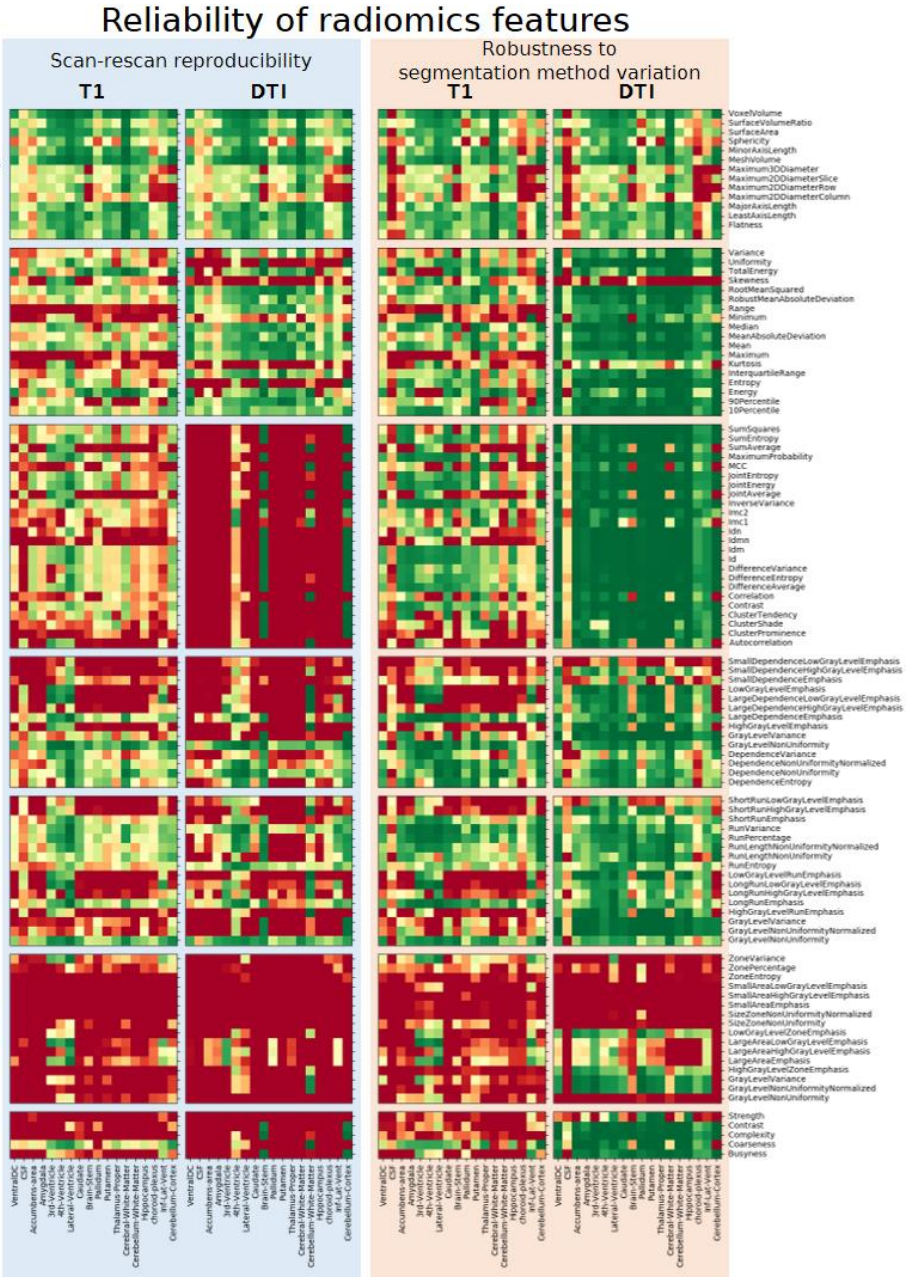


FreeSurfer SAMSEG **VUNO Med-DeepBrain** **FreeSurfer ASEG** **FastSurfer**

Two examples of the subtlety of differences in segmentation results for four different segmentation methods best seen in color print. From left to right: SAMSEG, Med-DeepBrain, FreeSurfer ASEG and FastSurfer. Subcortical differences are best visible lateral to the thalamus at the putamen (pink) and pallidum (blue), or cranial to the thalamus at the caudate nucleus (cyan) under the lateral ventricle (purple).

The yellow arrow points at label color for CSF. Only SAMSEG segments CSF surrounding the cortex, resulting in the low dice pair agreement seen earlier (Fig. 1).

Supplementary Figure S4. An extensive overview of the robustness of radiomic features.



Radiomic feature reliability measured using ICC. The left two columns show scan-rescan reproducibility. The right two show robustness to segmentation method variation. The color scale runs from red (worse) through yellow to green (better) with limits [0.5, 1].

The vertical axis is comprised of 107 radiomic features, categorized by feature class shown on the left side of the chart. Abbreviations of feature classes can be found in *Supplement B.* Vertical bar-patterns can be interpreted as anatomical-specific robustness deviation. A clear example of under-performing regions are the choroid plexus and inferior lateral ventricle in DTI segmentation method robust-ness.

On the horizontal axis of each column the same eighteen anatomical labels are showed as used in Figure 1.Horizontal bar-patterns can be interpreted as radiomic feature-specific robustness deviation. Take for example NGTDM-coarseness, scoring better than other NGTDM-features across the board.

The bar chart of Figure 2 shows an aggregation of each block of each column.

Supplementary Note S3. Description of segmentation methods

- ASEG is the most extensively used whole-brain segmentation method. ASEG uses a probabilistic-based atlas which captures information about the location of structures to determine the relationship between imagine intensity and neuroanatomical labels.[54] This type of atlas makes use of information regarding the statistical properties of anatomical structures to store the image a space in which coordinates have anatomical meaning as opposed to the somewhat arbitrary coordinates in a raw image. It is distributed via the FreeSurfer analysis package.[64] In this work, version 7.1.0 was used.
- SAMSEG is a recently introduced iteration on ASEGs atlas-based approach. SAMSEG uses generative parametric models which encodes probabilistic atlases using Bayesian interference.[97] SAMSEG is less computationally demanding and has shown improvement in robustness and sensitivity over FreeSurfer.[54]

The remaining two methods are deep learning-based and have been recently introduced.

- FastSurfer was introduced in 2020 and is based on a DenseNet architecture.[55]
- VUNO Med-DeepBrain is based on work by Suh et al., published in 2019. Their architecture is HighResNet-based.[56]

Both demonstrate high robustness on a scan-rescan dataset, and sensitivity on an Alzheimer's disease dataset. Only FastSurfer shows direct comparison to FreeSurfer as alternative method.

Supplementary Note S2. Description of radiomic feature classes

- Shape-based features,[3,84] also known as semantic features include volume, elongation and surface area.
- First-order features are often histogram-based and examine gray level signal intensity. Examples of these features include uniformity, entropy, mean, median and kurtosis.[1,84]
- Second-order features can be thought of as "texture" features. These examine spatial relationship between gray level signal intensities by constructing a gray-level dependence matrix[1,84] Three examples of types of features are:
    - Gray level co-occurrence matrix (GLCM) which analyzes occurrence of different gray level voxel pairs in different directions.[85]
    - Gray level dependence matrix (GLDM), with dependency defined as a [86]
    - Gray level run length matrix (GLRLM), which quantifies the number of consecutive voxels with same gray level. [87]
    - Gray level size zone matrix (GLSZM), which quantifies the size of a homogenous area of an image.[88]
    - Neighboring gray tone difference matrix (NGTDM), which quantifies the difference between a gray value and the average gray value of its neighbors.
- Higher-order features involve application of a filter or transformation to an image prior to feature extraction. These features aim to identify patterns or highlight details within the image that are hard to interpret and do not have a standardized definition.[98]